

1. The generic protein stiffness potential and secondary structure bias

5

As defined above, the model chain was intrinsically very flexible. A substantial fraction of its conformations that were allowed due to the assumed simplified hard core interactions did not correspond to any real polypeptide chain conformation. In particular, proteins are relatively stiff polymers. Moreover, folded proteins have very characteristic distributions of certain short-range distances. For example, the bimodal distribution of the distances between the i -th and $i+4^{\text{th}}$ residues reflects the tendency to adopt either of two types of conformations. These correspond to extended (β -type or extended coil) or very compact conformations (as within helices or turns). Such generic features need to be included in the model. Here, the SICHO model differs from that used in Example 1 due to the refined protein representation (a larger number of allowed chain vectors and a modified position of the center of interaction, that also included alpha carbons).

10

15

20

First, for all possible two-vector sequences of the model chain, a direction w was defined that was almost perpendicular to the plane formed by the fragment. A small systematic deviation from the exactly orthogonal direction was introduced in w to obtain vectors that were, on average, parallel to the helix axis and which also accounted for the average supertwist of β -strands.

$$u_i = (v_{i-1} \otimes v_i - v_{i-1} \cdot v_i) \quad (1)$$

$$w_i = u_i / |u_i| \quad (2)$$

25

30

where v_i is the i -th vector (or virtual bond) of the model chain, the symbol " \otimes " denotes the vector cross product and $|u_i|$ is the length of vector u_i . Consequently, these "directions of secondary structure" (the vectors w point along a helix or across a β -sheet) were normalized so that their length was equal to 1. The idea is explained in Figure 12, where the model chain virtual bonds are shown in solid lines and the vectors w_i are shown in open arrows.

The stiffness/secondary structure bias has the following form:

$$E_{\text{stiff}} = -\epsilon_{\text{gen}}[\sum \min\{0.5, \max(0, w_i \cdot w_{i-4})\}] \quad (3)$$

$$-\epsilon_{\text{gen}}[\Sigma \min\{0.5, \max(0, \mathbf{w}_i \cdot \mathbf{w}_{i-4})\}]$$

5 ϵ_{gen} is a constant energy parameter, common for all generic potentials, and Σ means summation along the chain. The above formulation means that the system is energetically stabilized when pairs of “direction of secondary structure” vectors are parallel (positive dot product). As can be read from the above equation, the stabilization energy increased in the range between 90° and 30° (angle between appropriate vectors \mathbf{w}) and then maintained its extreme value. Thus, small fluctuations of the secondary structure had no influence on the value of this potential, nor did the changes outside the regular conformations (negative values of the dot product) have any effect on the conformational energy. The minimum value of the stiffness function per residue was equal to ϵ_{gen} , and the maximum was 0.

15 Additionally, a bias was introduced towards the specific geometry of helical and β -type expanded states (however, it was quite permissively defined). All conformations were, of course, allowed; the purpose of this bias was to mimic a protein-like (average) distribution of local conformations. Symbolically, this could be written as follows:

$$20 \quad E_{\text{struct}} = \Sigma \{ \delta H1(i) + \delta H2(i) + \delta E1(i) + \delta E2(i) \} \quad (4)$$

with:

$$\begin{aligned} \delta H1(i) &= -\epsilon_{\text{gen}}, & \text{for } r_{i,i+4}^2 < 36 \text{ and } (\mathbf{v}_i \bullet \mathbf{v}_{i+3}) > 0 \text{ and } (\mathbf{v}_i \bullet \mathbf{v}_{i+2}) < -5 \\ &0, & \text{otherwise} \end{aligned} \quad (4a)$$

$$\begin{aligned} \delta H2(i) &= -\epsilon_{\text{gen}}, & \text{for } r_{i,i+4}^2 < 36 \text{ and } (\mathbf{v}_i \bullet \mathbf{v}_{i+3}) > 0 \text{ and } (\mathbf{v}_{i+1} \bullet \mathbf{v}_{i+3}) < -5 \\ &0, & \text{otherwise} \end{aligned} \quad (4b)$$

$$\begin{aligned} \delta E1(i) &= -\epsilon_{\text{gen}}, & \text{for } 56 < r_{i,i+4}^2 < 135 \text{ and } (\mathbf{v}_i \bullet \mathbf{v}_{i+2}) > 5 \\ &0, & \text{otherwise} \end{aligned} \quad (4c)$$

$$\begin{aligned} \delta E2(i) &= -\epsilon_{\text{gen}}, & \text{for } 56 < r_{i,i+4}^2 < 135 \text{ and } (\mathbf{v}_{i+1} \bullet \mathbf{v}_{i+3}) > 5 \\ &0, & \text{otherwise} \end{aligned} \quad (4d)$$

30 The numerical values are in lattice units and were selected to define a broad range of helical/turn conformations (for the $\delta H1$ and $\delta H2$ contributions) or expanded

conformations (for the $\delta E1$ and $\delta E2$ contributions). Due to the exclusive character
 5 of the two subsets of geometrical conditions for specific chain conformations, the
 minimum contribution from a residue is equal to $-2\epsilon_{\text{gen}}$ (either the first two
 conditions or the two last conditions can be simultaneously satisfied). Expressed
 differently, equation (4d) indicated that the system gained an energy equal to $-\epsilon_{\text{gen}}$
 for being in an expanded β -type conformation. For a four-vector fragment of the
 10 chain, this required that the distance between the i -th and $i+4^{\text{th}}$ beads (the centers of
 mass of the side chain plus $C\alpha$ units) had to lie between 10.7 and 16.8 Å, and the
 chain vectors v_{i+1} and v_{i+3} have to be oriented in a parallel-like fashion (the dot
 product >5). Additional stabilization is gained when, for the same fragment, another
 pair of vectors is parallel (*see* equation 4c). The broad ranges allowed for
 15 substantial fluctuations (without an energetic penalty) around an ideal expanded
 state and accommodated the variations of the model chain geometry caused by
 differences in side chain size.

Computational experiments have also been performed where all interactions,
 except the ones defined above, were turned off. At low temperature, the model
 20 chain formed rapidly fluctuating local clusters of expanded and helix-like states.
 The persistence length and the distributions of the short-range distances along the
 chains mimicked protein-like geometry.

2. Generic packing cooperativity

25 Two terms were introduced to enforce some of the most general regularities
 of the dense packing of protein structures.¹⁰ In all the more regular elements of
 secondary structure (within helices and β -sheets, but not between helices) and, to a
 lesser extent, in some coil-type fragments and turns, given a contact between a pair
 of reference residues, there was a very strong preference to have contacts (we
 30 provide precise definition of the “contact” later) between the preceding and the
 following residues. Indeed, the contact maps of globular proteins contain very
 characteristic strips.²² Those near the diagonal corresponded to the intrahelical